| | |
|---|---|
| REPORT DOCUMENTATION PAGE | Form Approved OMB NO. 0704-0188 |

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggesstions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any oenalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

| | | |
|---|---|---|
| 1. REPORT DATE (DD-MM-YYYY)<br>12-08-2014 | 2. REPORT TYPE<br>Final Report | 3. DATES COVERED (From - To)<br>12-May-2009 - 11-May-2016 |

| | |
|---|---|
| 4. TITLE AND SUBTITLE<br>Final Report: Taming Crowded Visual Scenes | 5a. CONTRACT NUMBER<br>W911NF-09-1-0255 |
| | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER<br>611102 |
| 6. AUTHORS<br>Mubarak Shah, Haroon Idrees | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Central Florida
12201 Research Parkway, Suite 501

Orlando, FL 32826 -3246

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
56014-CS.11

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

This report summarizes the progress made at Center for Research in Computer Vision at UCF in several distinct aspects of the project. We developed solutions to a number of important and challenging problems related to visual analysis of crowds and modeling of multi--agent interactions. The first category of solutions include behavioral analysis of crowds in videos captured through stationary as well as moving cameras. Second, we developed a novel technique to detect individuals in sparse crowds. This technique employs superpixels and iteratively improves the output of a generic underlying human detector. An advantage of the approach is that it outputs the exact

15. SUBJECT TERMS
Final Report, Crowded Scene Analysis, Anomaly Detection, Computer Vision

| | | | | | |
|---|---|---|---|---|---|
| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 15. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON<br>Mubarak Shah |
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | |
| UU | UU | UU | UU | | 19b. TELEPHONE NUMBER<br>407-823-5077 |

## Report Title

Final Report: Taming Crowded Visual Scenes

## ABSTRACT

This report summarizes the progress made at Center for Research in Computer Vision at UCF in several distinct aspects of the project. We developed solutions to a number of important and challenging problems related to visual analysis of crowds and modeling of multi--agent interactions. The first category of solutions include behavioral analysis of crowds in videos captured through stationary as well as moving cameras. Second, we developed a novel technique to detect individuals in sparse crowds. This technique employs superpixels and iteratively improves the output of a generic underlying human detector. An advantage of the approach is that it outputs the exact segmentation of humans besides the more common bounding boxes.

Third, we introduced and evaluated two new methods for analysis of extremely dense crowded scenes. The first approach deals with tracking of individuals in videos of high density crowd such as those depicting a marathon, while the other approach counts the number of individuals in an image of dense crowd. Finally, we developed a method that can detect anomalous behaviors in crowds using trajectories obtained through particle advection.

Besides that, we also published the cover article in the prestigious Communication of ACM (CACM) and a book on modeling and analysis of crowds by Springer. All of research was conducted under the ARO's support during this project.

The rest of the report describes these approaches in detail. There are five main sections: analysis of crowd behaviors, human detection in sparse crowds, visual analysis of extremely dense crowds, abnormal event detection, and finally we present details on the article and book published on visual analysis of crowds.

---

## Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

### (a) Papers published in peer-reviewed journals (N/A for none)

| Received | | Paper |
|---|---|---|
| 08/19/2010 | 1.00 | Shandong Wu, Brian E. Moore, Mubarak Shah. Chaotic Invariants of Lagrangian Particle Trajectories for Anomaly Detection in Crowded Scenes,<br>Twenty-Third IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (08 2010): . doi: |
| 08/19/2010 | 2.00 | Ramin Mehran, Brian E. Moore, Mubarak Shah. A Streakline Representation of Flow in Crowded Scenes,<br>Proceedings of the European Conference on Computer Vision (ECCV), (08 2010): . doi: |
| 08/31/2012 | 7.00 | Berkan Solmaz, Brian E. Moore, Mubarak Shah. Identifying Behaviors in Crowd Scenes Using Stability Analysis for Dynamical Systems,<br>IEEE Transactions on Pattern Analysis and Machine Intelligence, (10 2012): 2064. doi: 10.1109/TPAMI.2012.123 |
| 09/05/2013 | 9.00 | S. Khokhar, I. Saleemi, M. Shah. Multi-agent event recognition by preservation of spatiotemporal relationships between probabilistic models,<br>Image and Vision Computing, (09 2013): 0. doi: 10.1016/j.imavis.2013.06.004 |
| **TOTAL:** | | **4** |

**Number of Papers published in peer-reviewed journals:**

## (b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

**TOTAL:**

**Number of Papers published in non peer-reviewed journals:**

## (c) Presentations

**Number of Presentations:** 0.00

## Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

**TOTAL:**

**Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):**

## Peer-Reviewed Conference Proceeding publications (other than abstracts):

| Received | | Paper |
|---|---|---|
| 09/05/2013 | 10.00 | Afshin Dehghan, Mubarak Shah, Guang Shu. Improving an Object Detector and Extracting Regions using Superpixels,<br>Computer Vision and Pattern Recognition 2013. 24-JUN-13, . : , |
| **TOTAL:** | | **1** |

**Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):**

---

## (d) Manuscripts

| Received | | Paper |
|---|---|---|
| 04/05/2012 | 6.00 | Mubarak Shah, Brian E. Moore, Ramin Mehran, Saad Ali. Visual crowd surveillance through a hydrodynamics lens,<br>Communications of the ACM (12 2011) |
| 08/19/2010 | 3.00 | Berkan Solmaz, Brian E. Moore, and Mubarak Shah. Identifying Behaviors in Crowded Scenes Using Stability Analysis for Dynamical Systems,<br>(08 2010) |
| 11/14/2011 | 4.00 | Brian Moore, Saad Ali, Ramin Mehran, Mubarak Shah. Visual Crowd Surveillance Through A Hydrodynamic Lens,<br>Communications of the Association for Computing Machinery (11 2011) |
| 11/14/2011 | 5.00 | Haroon Idrees, Nolan Warner, Mubarak Shah. Tracking in Dense Crowds Using Prominenceand Neighborhood Motion Concurrence,<br>IEEE Transactions on Pattern Analysis and Machine Intelligence (11 2011) |
| **TOTAL:** | | **4** |

**Number of Manuscripts:**

---

## Books

| Received | Book |
|---|---|
| **TOTAL:** | |

| Received | | Book Chapter |
|---|---|---|
| 09/06/2013 | 8.00 | Ko Nishino, Dinesh Manocha, Saad Ali, Mubarak Shah. Modeling, Simulation and Visual Analysis of Crowds: A Multidisciplinary Perspective, Not available: Springer, (11 2013) |

**TOTAL:** 1

## Patents Submitted

MULTI-SOURCE, MULTI-SCALE COUNTING IN DENSE CROWD IMAGES filed on June 25, 2014 with U.S. Letters Patent as Serial No. 14/315,058

## Patents Awarded

## Awards

## Graduate Students

| NAME | PERCENT_SUPPORTED |
|---|---|

**FTE Equivalent:**
**Total Number:**

## Names of Post Doctorates

| NAME | PERCENT_SUPPORTED |
|---|---|

**FTE Equivalent:**
**Total Number:**

## Names of Faculty Supported

| NAME | PERCENT_SUPPORTED |
|---|---|

**FTE Equivalent:**
**Total Number:**

## Names of Under Graduate students supported

| NAME | PERCENT_SUPPORTED |
|---|---|

**FTE Equivalent:**
**Total Number:**

## Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: ...... 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:...... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):...... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense ...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:...... 0.00

## Names of Personnel receiving masters degrees

NAME

**Total Number:**

## Names of personnel receiving PHDs

NAME

**Total Number:**

## Names of other research staff

NAME PERCENT_SUPPORTED

**FTE Equivalent:**

**Total Number:**

## Sub Contractors (DD882)

## Inventions (DD882)

## Scientific Progress

See Attachment

## Technology Transfer

# Taming Crowded Visual Scenes

Final Project Report 2009-14

PI: Prof. Mubarak Shah

Center for Research in Computer Vision
University of Central Florida
Orlando FL 32826

Funding Agency:

U.S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709

# Scientific Progress and Accomplishments

This report summarizes the progress made at Center for Research in Computer Vision at UCF in several distinct aspects of the project. We developed solutions to a number of important and challenging problems related to visual analysis of crowds and modeling of multi-agent interactions. The first category of solutions include behavioral analysis of crowds in videos captured through stationary as well as moving cameras. Second, we developed a novel technique to detect individuals in sparse crowds. This technique employs superpixels and iteratively improves the output of a generic underlying human detector. An advantage of the approach is that it outputs the exact segmentation of humans besides the more common bounding boxes.

Third, we introduced and evaluated two new methods for analysis of extremely dense crowded scenes. The first approach deals with tracking of individuals in videos of high density crowd such as those depicting a marathon, while the other approach counts the number of individuals in an image of dense crowd. Finally, we developed a method that can detect anomalous behaviors in crowds using trajectories obtained through particle advection.

Besides that, we also published the cover article in the prestigious *Communication* of ACM (CACM) and a book on modeling and analysis of crowds by Springer. All of research was conducted under the ARO's support during this project.

The rest of the report describes these approaches in detail. There are five main sections: analysis of crowd behaviors, human detection in sparse crowds, visual analysis of extremely dense crowds, abnormal event detection, and finally we present details on the article and book published on visual analysis of crowds.

## 1. Analysis of crowd behaviors

Crowded scene analysis poses a challenging problem in the computer vision community. High densities of pedestrians in real-world situations make the recognition and tracking of individuals impractical. Nevertheless, it is important and meaningful to study this problem as automated detection of crowd behaviors, in particular, specific behavior patterns, has numerous applications. Some examples include prediction of congested areas, which may help avoid unnecessary crowding or clogging, and discovery of any abnormal events to help locate sudden changes in the behaviors of a crowd scene, etc.

### a. Crowd behaviors in stationary cameras

In this work, we developed a method for identifying crowd behaviors, e.g., bottlenecks, fountainheads, lanes, arches, and blocking, etc., in crowded scenes. These behaviors are illustrated visually in Fig. 1 below. In the proposed algorithm, a scene is overlaid by a grid of particles initializing a dynamical system defined by the optical flow. Time integration of the dynamical system then provides particle trajectories

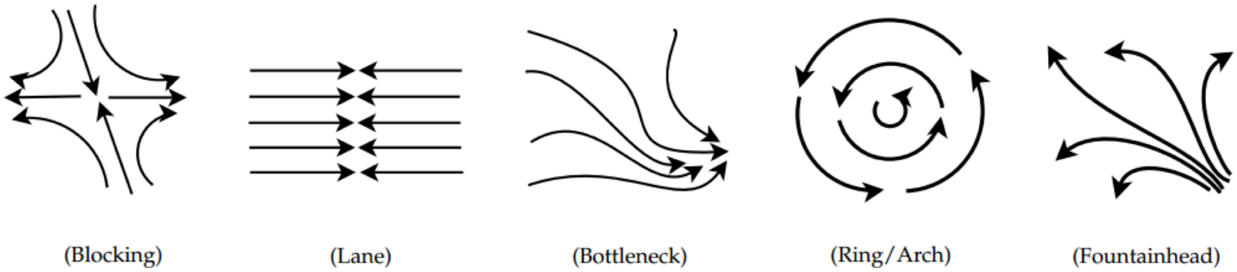


Figure 1: Diagrams showing 5 different crowd behaviors where the direction of motion is depicted by arrows.

that represent the motion in the scene. These trajectories are then used to locate regions of interest in the scene.

Linear approximation of the dynamical system provides behavior classification through the Jacobian matrix, where its Eigenvalues determine the dynamic stability of points in the flow and each type of stability corresponds to one of the five crowd behaviors. The Eigenvalues are only considered in the regions of interest, consistent with the linear approximation and the implicated behaviors. The algorithm is repeated over sequential clips of a video in order to record changes in Eigenvalues, which may imply changes in behavior. The overall process flow of the proposed approach is shown in Fig. 2.

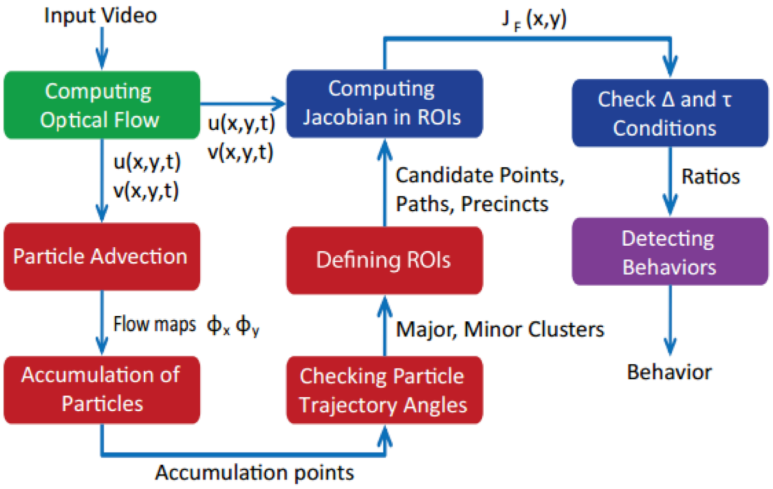


Figure 1: Overview of the proposed framework for crowd behaviors identification.

The method was tested on over 60 crowd and traffic videos. Some of the qualitative results are shown in Fig. 6. The videos were collected from Getty images, BBC motion gallery, YouTube, Thought Equity, and one sequences from the PETS 2009 dataset. The test videos were manually labeled to generate groundtruth behaviors. Following the PASCAL VOC challenge, detection accuracy was based on the overlap between automatically detected and groundtruth regions. An overlap of 40% was required for lanes and arches, and a distance of 40 pixels was used for the three other behaviors.

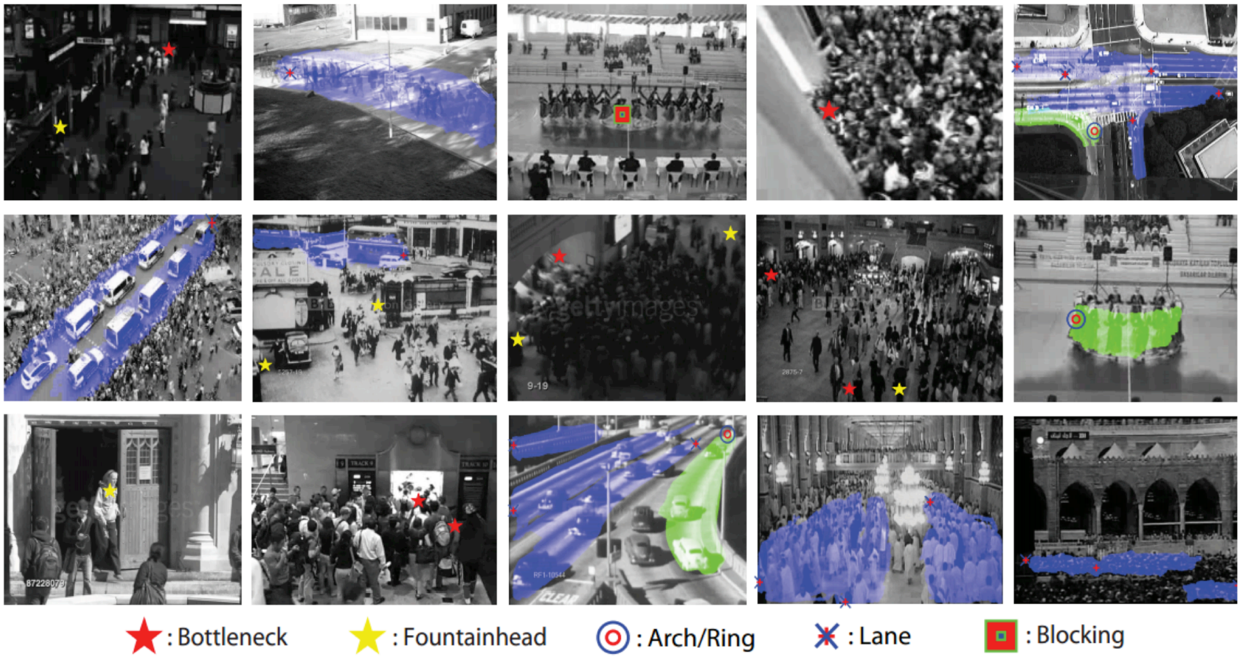


Figure 3: Scenes from 15 real video sequences, each showing behaviors detected automatically using our approach, as per the legend on the bottom.

Quantitative analysis of our experiments is shown in Table 1 below in terms of correct detections, false positive detections, and missed detections.

| Behavior | Total # of behaviors | # of Total Detections | # of Missed Detections | # of False Detections |
|---|---|---|---|---|
| **Lane** | 66 | 56 | 10 | 11 |
| **Blocking** | 3 | 3 | 0 | 0 |
| **Bottleneck** | 20 | 16 | 4 | 3 |
| **Fountainhead** | 29 | 23 | 7 | 5 |
| **Arch/Ring** | 28 | 23 | 5 | 6 |

Table 1: Quantitative analysis of proposed approach

The proposed framework has been experimentally verified to identify multiple crowd behaviors through stability analysis for dynamical systems, without the need of object detection, tracking or training. The promising results obtained demonstrate the capability and flexibility of the method for a wide variety of scenes. Despite its strengths, the method does have a few shortcomings: our model is deterministic and cannot capture the randomness inherent in the problem without a stochastic component; our model can only identify five behaviors, which is an oversimplification of the complexities encountered in crowds; and our method is not useful when a significant overlap of motion patterns is present in the scene, or when there is lack of consistent characteristic flow. These are open ended research questions and further work to alleviate the limitations of this method are being pursued. This research was accepted for publication in the IEEE Transactions on Pattern Analysis and Machine Intelligence in 2010.

## b. Crowd behaviors in moving cameras

This method tackles the problem of crowd behavior recognition in moving cameras videos for the first time. We have developed a framework for crowd behavior analysis in moving camera using Lie algebra. The method involves a novel motion segmentation algorithm adapted for the purpose of detecting regions of coherent motion in a video sequence of crowds. Even though this framework is proposed for crowd videos, its application is not limited and it can readily be applied to other application domains. The proposed algorithm has achieved very promising results on a challenging set of real life videos of crowds, where no existing method is applicable.

The overall framework is described in the following flowchart:

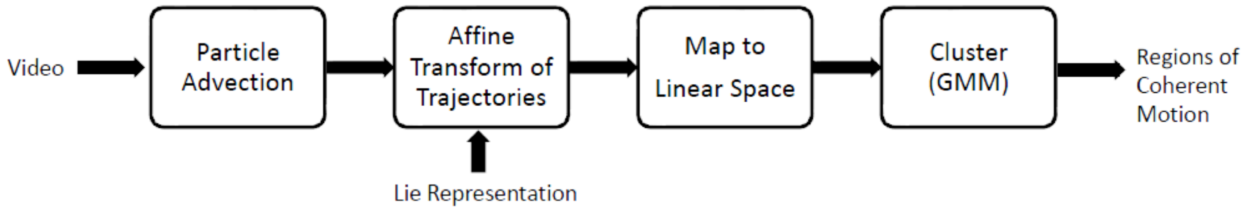


Figure 4: A high level flowchart describing our approach for crowd behavior recognition.

In the first step, we map dense particle trajectories from $(x, y)$ coordinates to a sequence of affine transforms. Specifically, we compute dense optical flow and then we perform particle advection to produce trajectories using 4th order Runga-Kutta integration. Next, we place a window of interest of size $n \times n$ over each pixel, which contains the patch of initial particle locations. The figure below illustrates a patch of particles for an example set of particle trajectories. The particles are seeded at the yellow rectangles and are moved using the frame-by-frame optical flow. The first patch of particles are illustrated as a blue rectangle. The second patch of particles contains the particles after one step of particle advection, and the third path contains the particles after second step of particle advection and so on.

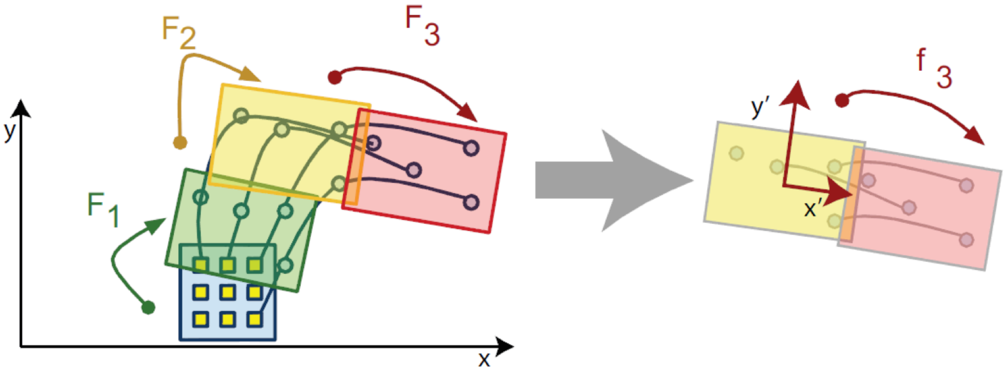


Figure 5: An illustration of a patch over pixels for computation of a sequence of affine transforms from particle trajectories. The seeding particles (Small yellow squares), particles after advection (small circles), the first particle patch (blue rectangle), particle patches after advection (large rectangles).

We use an affine transformation to model particle positions from one patch to another. Therefore, the trajectories of the window of interest are transformed into a sequence of affine transformation matrices, thereby capturing contextual information via motion of local patches, as compared to motion of individual particles.

The regions of coherent motion contain smooth sequences of affine transformations. Thus, segmentation of a video sequence reduces to finding regions with similar sequences of affine motion. However, the affine space lies on a nonlinear manifold due to a variety of geometric constraints which come from the fact that the affine group has a multiplicative rather than additive structure. This makes it difficult to perform segmentation of regions in affine space using clustering mechanisms. We therefore leverage Lie Algebra to map the affine transforms to a vector space which has an additive structure.

Given the parameterization of the particle trajectories as sequence of affine transformations in a vector space, we then learn a statistical model of motion at each frame. Using this model, we are able to cluster the motion in the video sequence into multiple rigid motions. The clusters are then used to initialize a Gaussian Mixture Model (GMM) using them as the components of the mixture. Finally, the Expectation Maximization (EM) algorithm is used to learn the optimal parameters of the mixture given the observations. KL divergence between distinct mixture distributions is then used to define a similarity measure between representations of various behaviors. This similarity is then employed to perform behavior recognition, the results of which are reported in the table on the right. The results are shown in Table 2.

| Behavior | TP | FP | Total |
|---|---|---|---|
| **Turn** | 32 | 6 | 38 |
| **Lane (to Right)** | 8 | 0 | 8 |
| **Lane (to Right)** | 4 | 0 | 4 |
| **Lane (to Left)** | 8 | 0 | 8 |
| **Lane (to Left)** | 7 | 0 | 0 |
| **Lane (to Right)** | 14 | 0 | 14 |
| **Lane (to Left)** | 4 | 1 | 5 |
| **Lane (to Down)** | 9 | 3 | 12 |
| **Lane (to Up)** | 14 | 1 | 15 |
| **Lane (to Left)** | 25 | 0 | 25 |
| **Lane (to Right)** | 15 | 1 | 16 |
| **Lane (to Left)** | 12 | 1 | 13 |

Table 2: Results of the proposed behavior discovery approach for 60 random frames from the NGSIM dataset.

## 2. Human detection in sparse crowds

We have developed an approach to improve the detection performance of a generic detector when it is applied to a particular video. An overview of our framework is depicted in Fig. 6. The performance of offline-trained objects detectors are usually degraded in unconstrained video environments due to variant illuminations, backgrounds and camera viewpoints. Moreover, most object detectors are trained using Haar-like features or gradient features but ignore video specific features like consistent color patterns.

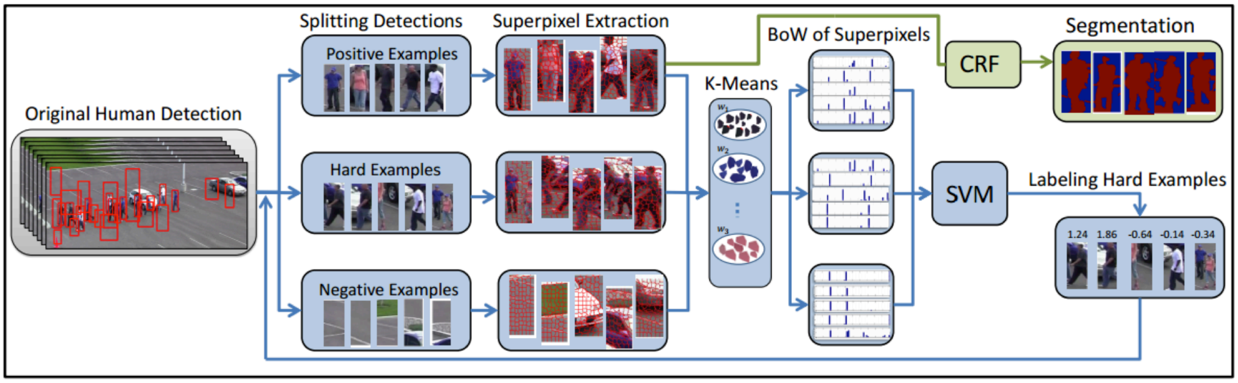
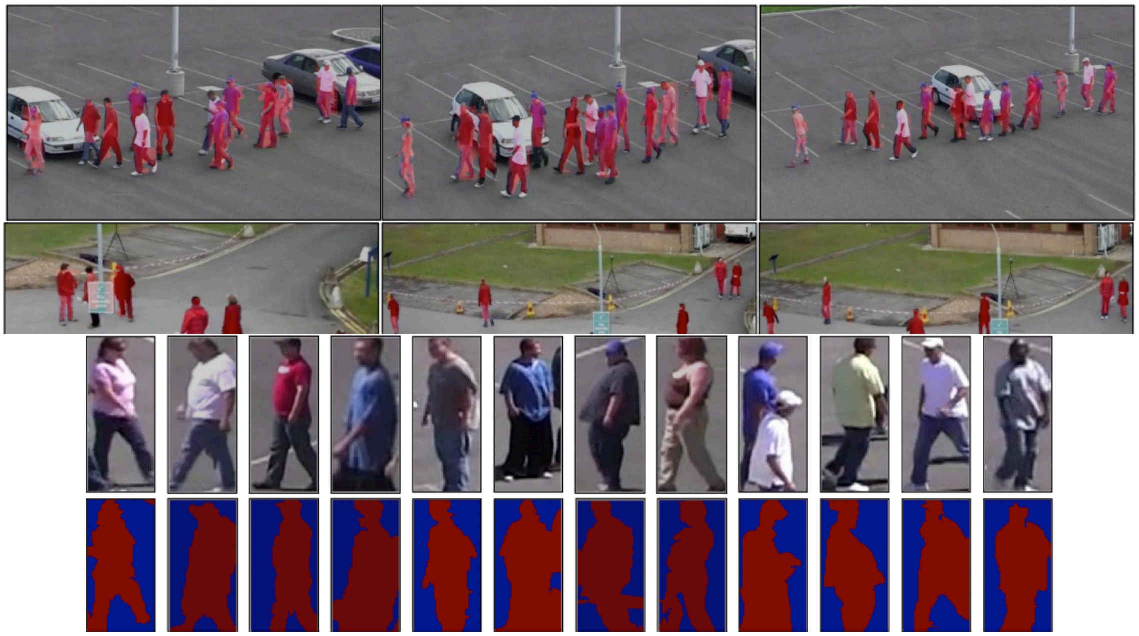


Figure 6: An overview of our approach for improving a generic human detector and segmentation of detected objects.

In our approach, we apply a Superpixel-based Bag-of-Words (BoW) model to iteratively refine the output of a generic detector. Compared to other related work, our method builds a video-specific detector using superpixels, hence it can handle the problem of appearance variation. Most importantly, using Conditional Random Field (CRF) along with our super pixel-based BoW model, we develop an algorithm to segment the object from the background. Therefore our method generates an output of the exact object regions instead of the bounding boxes generated by most detectors. In general, our method takes detection bounding boxes of a generic detector as input and generates the detection output with higher average precision and precise object regions. The experiments on four recent datasets demonstrate the effectiveness of our approach and significantly improves the state-of-art detector by 5-16% in average precision. The resutls are hown in Figure 6-7.



More technical details of the devised framework can be found in the research paper titled “Improving an Object Detector and Extracting Regions using Superpixels” published in IEEE CVPR 2013.

## 3. Visual analysis of extremely dense crowds

Visual analysis of dense crowds is particularly challenging due to large number of individuals, occlusions, clutter, and fewer pixels per person which rarely occur in ordinary surveillance scenarios. The solutions presented in this section aim to address these challenges in images and videos of extremely dense crowds containing hundreds to thousands of humans. In particular, we present solutions developed to track and count individuals in extremely dense crowds.

### a. Tracking in Dense Crowds

Methods designed for tracking in dense crowds typically employ *a priori* information to make this difficult problem tractable. In this work, we have shown that it is possible to handle this problem, without any priors, by taking into account the visual and contextual information already available in such scenes.

We have developed a novel tracking method tailored to dense crowds which does not require modeling of crowd flow and, at the same time, is less likely to fail in the case of dynamic crowd flows and anomalies by minimally relying on previous frames. To this end, we automatically identify prominent individuals from the crowd that are easier to track based on discriminative appearance. For example, as shown in the figure 9 on the right, some of the individuals in the very dense crowd stand out due to their appearance, and are marked by yellow bounding boxes.

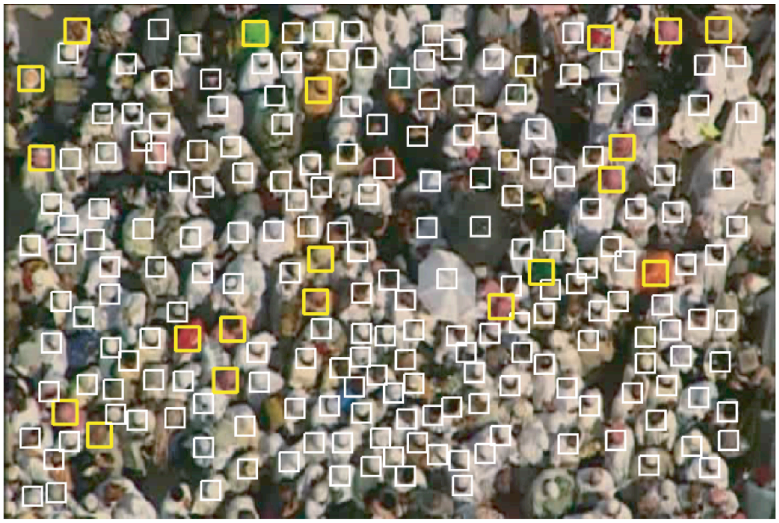


Figure 9: An example of a dense crowd where individuals highlighted in yellow squares stand out from the crowd and therefore should be easier to track.

In this work, we refer to these prominent individuals as “queens” (a reference to the queen bee in a colony of bees). In order to automatically choose the queens from among a very large number of individuals, we begin by collecting a 3d vector of RGB colors for each pixel in the individual bounding boxes of all individuals. These feature vectors are clustered into $k$ groups, which are then sorted by “cluster density”. The density of a cluster is defined as the ratio of mass and volume, which in our case correspond to the number of pixels in a cluster, and the volume of the ellipsoid representing the covariance of a cluster, respectively.

Given this ordering of clusters, the features (or pixels) in each cluster are assigned back to their respective templates (bounding boxes), starting from the first cluster. This process is stopped once 10% of total number of templates are at least two-thirds filled, which are then chosen as the queens. The underlying idea is that the first cluster in the list being the most dense, contains the most discriminative features across the feature space, and the templates which contributed to this cluster are likely to have appearance which is different from most other individuals.

Next, we use Neighborhood Motion Concurrence to model the behavior of individuals in a dense crowd which predicts the position of an individual based on the motion of its neighbors. The motion concurrence can be considered to be an additional cue used in data association, where instead of the object's motion model, we proposed to use the velocity vectors from the neighboring objects that have already been updated for the frame under consideration. The main idea behind this approach is illustrated in the figure 10 below.

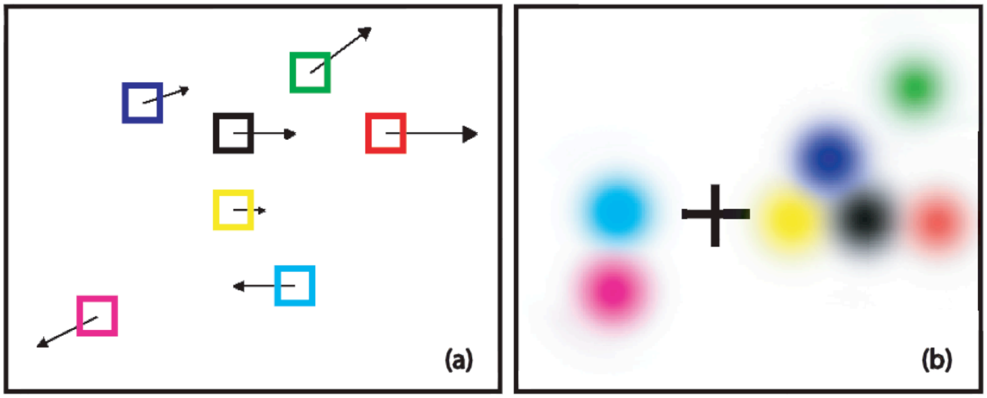


Figure 10: Visualization for Neighborhood Motion Concurrence model. (a) shows target in black whose position is to be updated and its updated neighbors shown in colors. (b) is the probability of position using the model for the target in (a). The cross hair represents position of the target before update.

Finally, we use a three-frame instantaneous flow to capture dynamic aspect of crowd behavior. When the individual moves with the crowd flow, we use Neighborhood Motion Concurrence to predict motion while leveraging instantaneous flow in case of dynamically changing flow and anomalies. Experiments on a number of sequences show that proposed solution can track individuals in dense crowds without requiring any pre-processing, making it a suitable online tracking algorithm. Some of these results along with experiments using competitive methods are shown in the figure 11 and Table 3 below. More results and experimental details can be seen in the research paper titled "Tracking in Dense Crowds

Using Prominence and Neighborhood Motion Concurrence" published in Image and Vision Computing 2014.

| | Seq 1 | Seq 2 | Seq 3 | Seq 4 | Seq 5 | Seq 6 | Seq 7 | Seq 8 |
|---|---|---|---|---|---|---|---|---|
| **# Frames** | 840 | 134 | 144 | 492 | 464 | 333 | 494 | 126 |
| **# People** | 152 | 235 | 175 | 747 | 171 | 600 | 73 | 58 |
| **Template Size** | 14 | 16 | 14 | 16 | 8 | 10 | 10 | 10 |
| **NCC** | 49% | 85% | 58% | 52% | 33% | 52% | 50% | 86% |
| **MS** | 19% | 67% | 16% | 8% | 7% | 36% | 28% | 43% |
| **MSBP** | 57% | 97% | 71% | 69% | 51% | 81% | **68%** | **94%** |
| **FF** | 74% | 99% | 83% | 88% | 66% | 90% | **68%** | 93% |
| **CTM** | 76% | **100%** | 88% | 92% | 72% | **94%** | 65% | **94%** |
| **Proposed** | **80%** | **100%** | **92%** | **94%** | **77%** | **94%** | 67% | 92% |

Table 3: Quantitative comparison between the proposed and five alternative methods where tracks that lie within 15 pixels of ground truth are considered to be correct. The bold numbers indicate the best performance for each sequence.

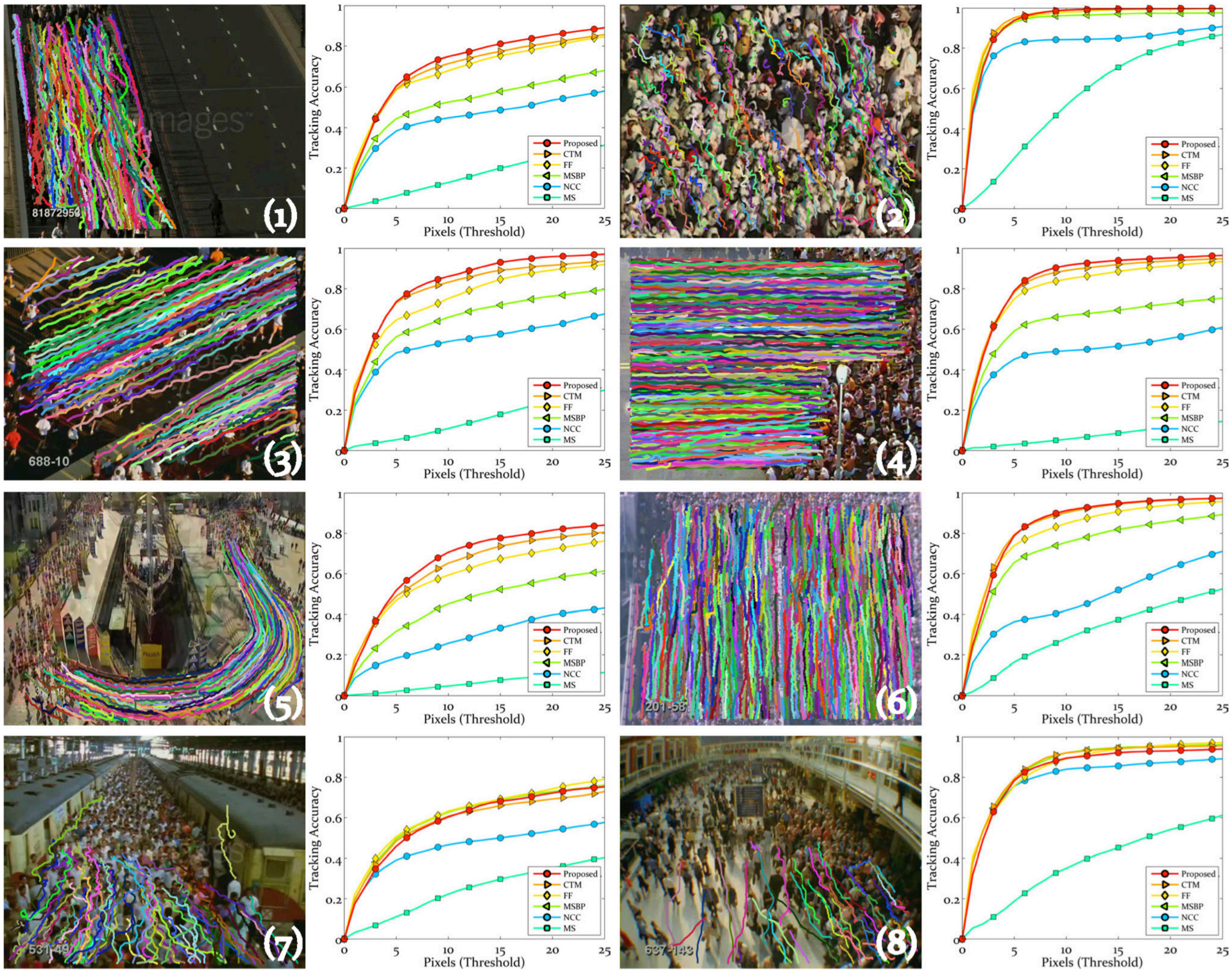


Figure 11: Results of the proposed method for individual tracking is shown for a few sequences. The plots show a quantitative comparison of our method (red) with some of the competitive tracking algorithms.

## b. Counting in Dense Crowds

For counting, we developed a framework that leverages multiple sources of information to estimate the exact number of individuals present in an extremely dense crowd visible in a single image. Our devised method overcomes problems including perspective, occlusion, clutter, and few pixels per person. Our approach relies on multiple sources including low confidence head detections, repetition of texture elements (using SIFT), and frequency-domain analysis to estimate counts, along with confidences associated with observing individuals, in an image region. Moreover we employ a global consistency constraint on counts using Markov Random Field. This caters for disparity in counts in local neighborhoods and across scales. Our proposed framework is applicable to images containing extremely dense crowds, ranging from one hundred to a few thousand humans.

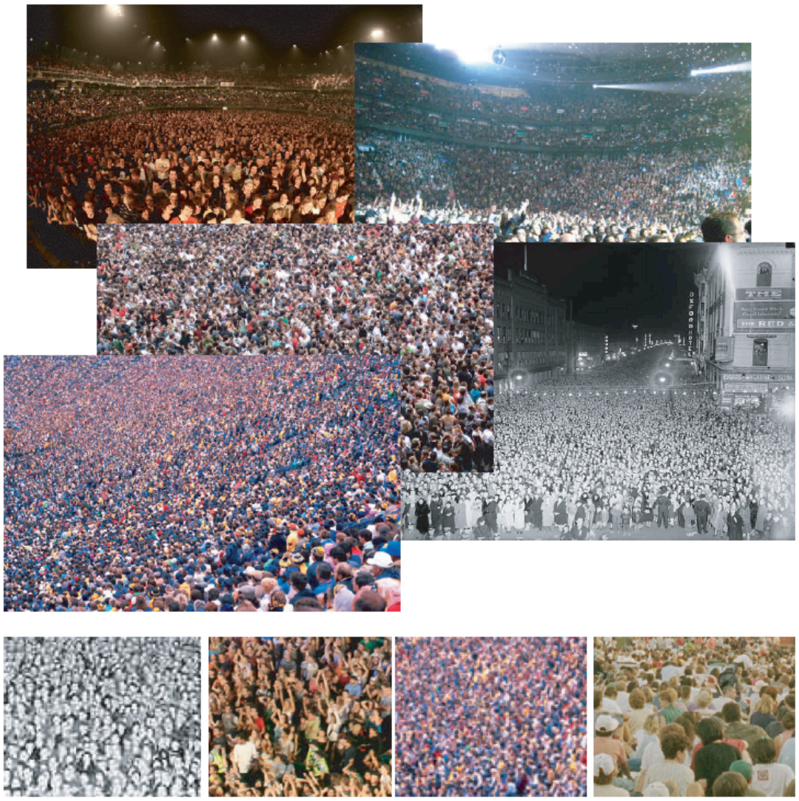


Figure 12: This figure shows five arbitrary images on which we tested the proposed framework. On average, each image in the tested crowd counting dataset contains around 1280 humans. The bottom row shows four patches from different images at the original resolution.

The problem of counting the number of objects, specifically people, in images and videos arises in several real world applications including crowd management, design and analysis of buildings and spaces, and safety and security. In certain scenarios, obtaining the people count is of direct importance, e.g., in public rallies, marathons, public parks, and transportation hubs, etc. The manual counting of individuals in very dense crowds is an extremely laborious task, but is performed nonetheless by experienced personnel when needed.

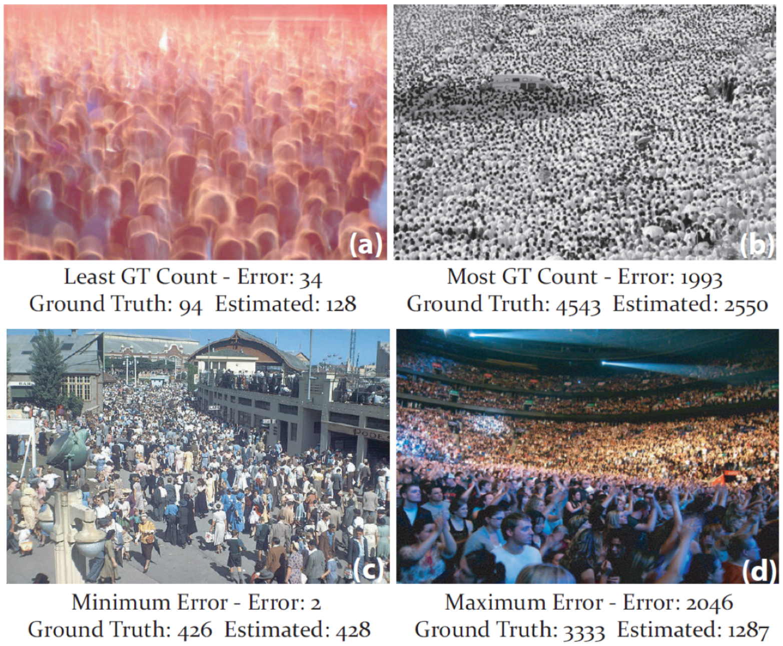
Computer vision research in the area of crowd analysis has resulted in several automated and semi-automated solutions for density estimation and counting. Practical application of most existing techniques however, is constrained by two important limitations: (1) inability to handle crowds of hundreds or thousands (Fig. 12) rather than a few tens of individuals; and (2) reliance on temporal constraints in crowd videos, which are not applicable to the more prevalent still images.

Most existing methods can be categorized by the application scenario and experimental setup. Some methods proposed in literature for crowd detection perform image segmentation without actual counting or localization, while others simply estimate the coarse density range within local regions. In terms of experimental data, most of the existing algorithms for exact counting have been tested on low to medium density crowds, ranging from a few to about 50 individuals per image. Our proposed technique has been tested on still images containing between 94 and 4543 people per image, with an average of 1,280 people over fifty images in the dataset. Such high density implies that an individual may occupy so few pixels that it can neither be detected, nor can its presence be verified given the location, which are key requirements in existing techniques.

In our proposed framework, we first employ Fourier along with head detections and interest point based counts in local neighborhoods on multiple scales to avoid the problem of irregularity in the perceived textures emanating from images of dense crowds. The count estimates from this localized multi-scale analysis are then aggregated subject to global consistency constraints. Secondly, in order to leverage multiple estimates from distinct sources, the corresponding confidence maps need to be comparable and in the same space. For instance, the Fourier transform is not directly useful in this regard, since it cannot be combined with count estimate maps in the image domain. We therefore reconstruct the low to medium frequency component of the image region and the reconstructed image is then compared with the original image after alignment. This process provides two important pieces of information: the estimated count per local region, and a measure of error relative to the original image.

Combining the three sources, i.e., Fourier, interest points, and head detection, with their respective confidences, we compute counts at localized patches independently, which are then globally constrained to get an estimate of count for the entire image. Since the data terms are evaluated independently at different scales, the smoothness constraint has to be applicable to spatial neighborhoods, as well as immediate neighbors at different scales. We devised a solution to obtain counts from multi-scale grid MRF which infers the solution simultaneously at all scales while enforcing the count consistency constraint.

Some of the results of the proposed approach are shown in Fig 13 More results, technical details, and discussion of the framework are presented in the



Least GT Count - Error: 34
Ground Truth: 94 Estimated: 128

Most GT Count - Error: 1993
Ground Truth: 4543 Estimated: 2550

Minimum Error - Error: 2
Ground Truth: 426 Estimated: 428

Maximum Error - Error: 2046
Ground Truth: 3333 Estimated: 1287

Figure 13: Selected images with their counts and errors with respect to ground truth. The first row (a, b) shows the extreme ends of the experimental dataset in terms of counts. The second row (c, d) shows the images with the lowest and the highest errors.

research paper titled “Multi-Source Multi-Scale Counting in Extremely Dense Crowd Images” published in IEEE CVPR 2013.

## 4. Abnormal event detection

In a crowded situation, it will be invaluable to have an automated mechanism for locating abnormal events in the videos of the crowded scene. Such a capability will give public safety officials enough time to take remedial actions. To solve the abnormal event detection problem, we will make use of the observation that the dynamics of a crowd flow segment can be depicted in terms of the particle trajectories.

Particle trajectories are generated by advecting the grid of particles through the flow field of the crowd. Each trajectory relates a particle’s initial position to its position at a later time, which represents the dynamical behavior of the crowd along its spatial extent. Hence, the collection of particle trajectories contains rich information about the current state/behavior of the crowd. This means that analyzing the particle trajectories in terms of dynamical properties serves as an effective means to investigate the crowded behaviors. It should be noted that using particle trajectories is advantageous in describing crowded scenes as it is actually applicable for and extractable from arbitrary crowded scenarios. In particular, considering that it is quite hard to represent a crowded scene of completely random behaviors using traditional object segmentation or tracking methods, particle trajectories can play an important role in such cases.

In order to effectively summarize the dynamical behavior of each particle trajectory, we further encode it in terms of the Chaotic Invariants, from which we can build a concise representation of crowded behaviors by calculating the features of metric and dynamical invariants. By examining how the chaotic feature set changes with respect to time in a crowd video, we will be able to find out the moment and spots of dramatic change of dynamic properties, which lead us to infer that an abnormal event occurs.

Next, we briefly describe the designed algorithm steps for locating abnormal events, as outlined in Fig. 14. For a video sequence of a crowd scene, optical flow will be first calculated. Then the video clip is temporally segmented to multiple blocks in terms of a predefined block size. The block size has a direct effect on the detection precision of the moments when abnormal events happen. For each block, we perform the particle advection, giving rise to the particle trajectories from frame one to frame ten (in a block). As the particle is treated in terms of a pixel, there will be a large number of particle trajectories. To make the trajectory set more compact, a simple trajectory clustering is applied by imposing a gridding-window sliding from the top left to bottom right of a frame/image. Here, the size of gridding-window is adjustable governing a coarse-to-fine representation manner. For example, we can choose 16*16 as the size of the window. After that, we can have multiple 3-D cuboids generated from the perspective of temporal segmenting (block) and spatial clustering (gridding-window). The trajectories in a cuboid are simply averaged to represent the motion in the cuboid, where we define the averaged trajectory as cuboid trajectory.

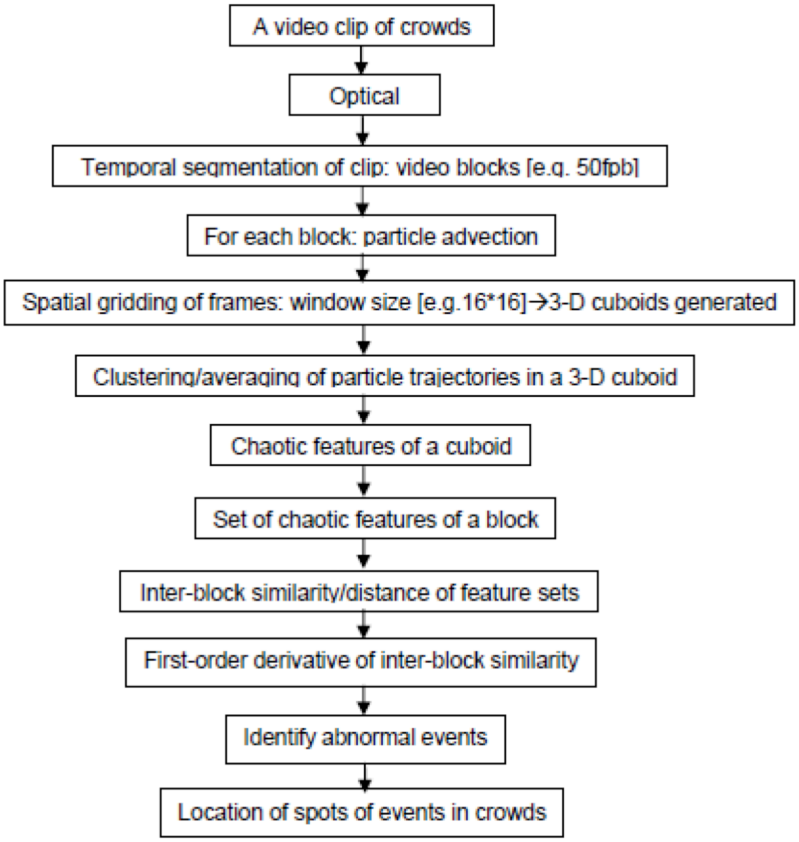

A video clip of crowds
Optical
Temporal segmentation of clip: video blocks [e.g. 50fpb]
For each block: particle advection
Spatial gridding of frames: window size [e.g.16*16]→3-D cuboids generated
Clustering/averaging of particle trajectories in a 3-D cuboid
Chaotic features of a cuboid
Set of chaotic features of a block
Inter-block similarity/distance of feature sets
First-order derivative of inter-block similarity
Identify abnormal events
Location of spots of events in crowds

Figure 14: Algorithmic steps designed for abnormal event detection

Subsequently, we calculate the chaotic features for a cuboid trajectory, which means we first embed the trajectory vectors into a phase space by finding optimal embedding parameters. One is the time-delay and the other is the embedding dimension. The former is achieved by mutual information maximization method and the false nearest neighbor algorithm is utilized for the latter. In total we calculate seven chaotic features: maximal Lyapunov exponent, ii) Correlation integral, iii) Correlation dimension, iv) Mean, v) Variance, vi) Time-delay, vii) Embedding dimension. Here, Lyapunov exponent is a dynamical invariant measuring the exponential divergence of the nearby trajectories in the phase space. The correlation integral is a metric invariant quantifying the density of points in the phase space. The correlation dimension also characterizes the metric structure and it measures the change in the density of phase space with respect to the neighborhood radius. Note that the same procedures for 7-D chaotic feature vectors are applied to the cuboid trajectories of all the cuboids. Stacking these chaotic features, we can have a chaotic feature set representing the chaotic dynamics of a block. Then we repeat the above steps for all of the blocks of the video clip

So far we are able to characterize the dynamics of all the blocks of the video. The next action will be to examine how the dynamical features change with respect to the temporal stamps (in terms of indexes of blocks). The normalized Euclidean distance of two consecutive feature sets will be computed to measure the dynamics similarity. Once we find a change greater than a predefined threshold by calculating the first-order derivative of inter-block similarity, we infer there is a dramatic change of crowd behavior at that moment, which could be an abnormal event worth paying extra attention or taking specific actions. The moments of abnormal events happening will be alerted and reported immediately.

For evaluation, we used the publicly available UMN dataset, which includes 11 video clips of 3 different scenarios of crowd escape panic. For each UMN clip, a normal starting section is followed by an abnormal ending section. The purpose of the experiment is to locate the moment where the abnormal event happens in the video clips. The block size is defined by 80fpb. Here we show the results on one video sequence in the following figures. Fig. 15 shows 9 frames of the sequence (in order of #1, #150, #320, #410, #485, #492, #539, #558, #590).

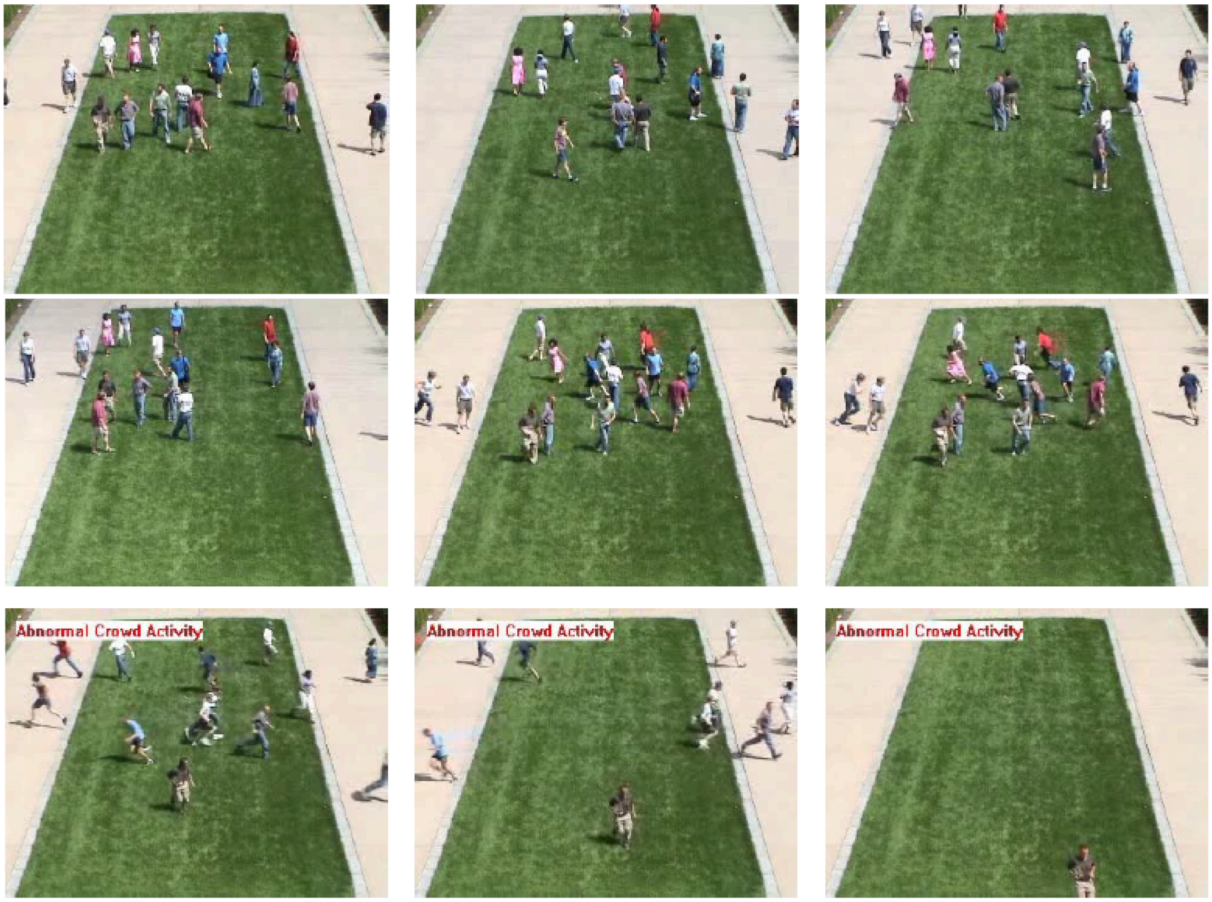


Figure 15: Nine frames of a video sequence of the UMA dataset. The bottom row shows frames where abnormal event was detected.

## 5. Other publications

### a. Modeling, Simulation and Visual Analysis of Crowds: A Multidisciplinary Perspective

This book recently published in the Springer International Series in Video Computing, discusses common challenges and points to problem areas related to modeling, simulation, and visual analysis of crowds. It facilitates the process of cross-disciplinary interaction among researchers from areas of computer vision, computer graphics and evacuation dynamics by providing a common platform, and provides a comprehensive map of the current state of the art in these distinct but related fields.

Over the last several years there has been a growing interest in developing computational methodologies for modeling and analyzing movements and behaviors of "crowds" of people. This interest spans several scientific areas that includes Computer Vision, Computer Graphics, and Pedestrian Evacuation Dynamics. Despite the fact that these different scientific fields are trying to model the same physical entity (i.e., a crowd of people), research ideas have evolved independently. As a result each discipline has developed techniques and perspectives that are characteristically their own.

The goal of this book to provide the readers a comprehensive map towards the common goal of better analyzing and synthesizing the pedestrian movement in dense, heterogeneous crowds. The monograph is organized into different parts that consolidate various aspects of research towards this common goal, namely the modeling, simulation, and visual analysis of crowds. Through this book, readers will see the common ideas and vision as well as the different challenges and techniques that will stimulate novel approaches to fully grasping crowds.

The book has been co-edited by Saad Ali, Mubarak Shah, Ko Nishino, and Dinesh Manocha.

## b. Visual Crowd Surveillance using Particle Dynamics

The research conducted at the UCF CRCV over the past years, especially the work sponsored by the Army Research Lab and Army Research Office has been novel, widely applicable, and theoretically interesting towards a variety of commonly encountered problems in visual analysis of crowded scenes. Part of this research was accepted for publication in the prestigious Communications of the ACM as a summary article titled "Visual Crowd Surveillance through a Hydrodynamics Lens".

The different methods and algorithms reported in this article include particle advection and particle trajectories for flow field generation, FTLE field generation, crowd behavior segmentation, abnormal behavior detection using social force modeling, and object tracking in extremely dense crowded environments. These methods were summarized for the general computer science audience and were extracted from several publications by different authors at the Center for Research in Computer Vision.

Figure 16: Cover page of the Communications of the ACM featuring a summarization of UCF CRCV's work in crowd analysis funded by ARO.